ENVIRONMENTAL HEALTH PERSPECTIVES

http://www.ehponline.org

CERAPP: Collaborative Estrogen Receptor Activity Prediction Project

Kamel Mansouri, Ahmed Abdelaziz, Aleksandra Rybacka, Alessandra Roncaglioni, Alexander Tropsha, Alexandre Varnek, Alexey Zakharov, Andrew Worth, Ann M. Richard, Christopher M. Grulke, Daniela Trisciuzzi, Denis Fourches, Dragos Horvath, Emilio Benfenati, Eugene Muratov, Eva Bay Wedebye, Francesca Grisoni, Giuseppe F. Mangiatordi, Giuseppina M. Incisivo, Huixiao Hong, Hui W. Ng, Igor V. Tetko, Ilya Balabin, Jayaram Kancherla, Jie Shen, Julien Burton, Marc Nicklaus, Matteo Cassotti, Nikolai G. Nikolov, Orazio Nicolotti, Patrik L. Andersson, Qingda Zang, Regina Politi, Richard D. Beger, Roberto Todeschini, Ruili Huang, Sherif Farag, Sine A. Rosenberg, Svetoslav Slavov, Xin Hu, and Richard S. Judson

http://dx.doi.org/10.1289/ehp.1510267

Received: 27 May 2015
Accepted: 8 February 2016
Advance Publication: 23 February 2016

CERAPP: Collaborative Estrogen Receptor Activity Prediction Project

Kamel Mansouri,[1*] Ahmed Abdelaziz,[2] Aleksandra Rybacka,[3] Alessandra Roncaglioni,[4] Alexander Tropsha,[5] Alexandre Varnek,[6] Alexey Zakharov,[7] Andrew Worth,[8] Ann M. Richard,[1] Christopher M. Grulke,[1] Daniela Trisciuzzi,[9] Denis Fourches,[5] Dragos Horvath,[6] Emilio Benfenati,[4] Eugene Muratov,[5] Eva Bay Wedebye,[10] Francesca Grisoni,[11] Giuseppe F. Mangiatordi,[9] Giuseppina M. Incisivo,[4] Huixiao Hong,[12] Hui W. Ng,[12] Igor V. Tetko,[2] Ilya Balabin,[13] Jayaram Kancherla,[1] Jie Shen,[14] Julien Burton,[8] Marc Nicklaus,[7] Matteo Cassotti,[11] Nikolai G. Nikolov,[10] Orazio Nicolotti,[9] Patrik L. Andersson,[3] Qingda Zang,[15] Regina Politi,[5] Richard D. Beger,[16] Roberto Todeschini,[11] Ruili Huang,[17] Sherif Farag,[5] Sine A. Rosenberg,[10] Svetoslav Slavov,[16] Xin Hu,[17] and Richard S. Judson[1]

[1]National Center for Computational Toxicology, U.S. Environmental Protection Agency, Research Triangle Park, North Carolina, USA; [2]Institute of Structural Biology, Helmholtz Zentrum Muenchen - German Research Center for Environmental Health (GmbH), Munich, Germany; [3]Chemistry Department, Umeå University, Umeå, Sweden; [4]Environmental Chemistry and Toxicology Laboratory, IRCCS-Istituto di Ricerche Farmacologiche Mario Negri, Milan, Italy; [5]Laboratory for Molecular Modeling, University of North Carolina, Chapel Hill, North Carolina, USA; [6]Laboratoire de Chemoinformatique, University of Strasbourg, Strasbourg, France; [7]National Cancer Institute, National Institutes of Health, Bethesda, Maryland, USA; [8]Institute for Health and Consumer Protection (IHCP), Joint Research Centre of the European Commission in Ispra, Ispra, Italy; [9]Department of Pharmacy-Drug Sciences, University of Bari, Bari, Italy; [10]National Food Institute, Division of Toxicology and Risk Assessment, Technical

University of Denmark, Copenhagen, Denmark; [11]Milano Chemometrics and QSAR Research Group, University of Milano-Bicocca, Milan, Italy; [12]National Center for Toxicological Research, Division of Bioinformatics and Biostatistics, U.S. Food and Drug Administration, Jefferson, Arizona, USA; [13]High Performance Computing, Lockheed Martin, Research Triangle Park, North Carolina, USA; [14]Research Institute for Fragrance Materials, Inc., Woodcliff Lake, New Jersey, USA; [15]Integrated Laboratory Systems, Inc., Research Triangle Park, North Carolina, USA; [16]National Center for Toxicological Research, Division of Systems Biology, U.S. Food and Drug Administration, Jefferson, Arizona, USA; [17]National Center for Advancing Translational Sciences, National Institutes of Health, Bethesda, Maryland, USA. *Additional affiliation: Oak Ridge Institute for Science and Education, Oak Ridge, Tennessee.

Address correspondence to Richard Judson, U.S. EPA, National Center for Computational Toxicology, 109 T.W. Alexander Drive, Research Triangle Park, NC 27711 USA

Telephone: 919-541-3085. E-mail: judson.richard@epa.gov

Running title: CERAPP

Acknowledgments: Ilya Balabin is employed by Lockheed Martin, RTP, NC, USA. Jie Shen is employed by Research Institute for Fragrance Materials, Inc., Woodcliff Lake, NJ, USA. Qingda Zang is employed by Integrated Laboratory Systems, Inc., RTP, NC, USA

Disclaimer: The views expressed in this paper are those of the authors and do not necessarily reflect the views or policies of the U.S. Environmental Protection Agency and the U.S. Food and Drug Administration.

Competing financial interests: The authors declare they have no actual or potential competing financial interests.

Abstract

Background: Humans are exposed to thousands of man-made chemicals in the environment. Some chemicals mimic natural endocrine hormones and, thus, have the potential to be endocrine disruptors. Most of these chemicals have never been tested for their ability to interact with the estrogen receptor (ER). Risk assessors need tools to prioritize chemicals for evaluation in costly *in vivo* tests, for instance, within the EPA Endocrine Disruptor Screening Program (EDSP).

Objectives: Here, we describe a large-scale modeling project called CERAPP (Collaborative Estrogen Receptor Activity Prediction Project) and demonstrate the efficacy of using predictive computational models trained on high-throughput screening data to evaluate thousands of chemicals for ER-related activity and prioritize them for further testing.

Methods: CERAPP combined multiple models developed in collaboration among 17 groups in the United States and Europe to predict ER activity of a common set of 32,464 chemical structures. Quantitative structure-activity relationship models and docking approaches were employed, mostly using a common training set of 1677 chemical structures provided by US EPA, to build a total of 40 categorical and 8 continuous models for binding, agonist, and antagonist ER activity. All predictions were evaluated on a set of 7,522 chemicals curated from the literature. To overcome the limitations of single models, a consensus was built by weighting models on scores based on their evaluated accuracies.

Results: Individual model scores ranged from 0.69 to 0.85, showing high prediction reliabilities. Out of the 32,464 chemicals, the consensus model predicted 4,001 chemicals (12.3%) as high priority actives and 6,742 potential actives (20.8%) to be considered for further testing.

Conclusion: This project demonstrated the possibility to screen large libraries of chemicals using a consensus of different *in silico* approaches. This concept will be applied in future projects related to other endpoints.

Introduction

There are tens of thousands of natural and synthetic chemical substances to which humans and wildlife are exposed (Dionisio et al. 2015; Egeghy et al. 2012; Judson et al. 2009). A subset of these compounds may disrupt normal functioning of the endocrine system and cause health hazards to both humans and ecological species (Birnbaum and Fenton 2003; Diamanti-Kandarakis et al. 2009; Mahoney and Padmanabhan 2010; UNEP and WHO 2013). Endocrine-disrupting chemicals (EDCs) can mimic or interfere with natural hormones and alter their mechanisms of action at the receptor level, as well as interfere with the synthesis, transport, and metabolism of endogenous hormones (Diamanti-Kandarakis et al. 2009). Exposure to EDCs can lead to adverse health effects involving developmental, neurological, reproductive, metabolic, cardiovascular, and immune systems in humans and wildlife (Colborn et al. 1993; Davis et al. 1993; Diamanti-Kandarakis et al. 2009).

The estrogen receptor (ER) is one of the most extensively studied targets related to the effects of EDCs (Mueller and Korach 2001; Shanle and Xu 2011). This concern about estrogen-like activity of man-made chemicals is because of their potential for negatively affecting reproductive function (Hileman 1994; Kavlock et al. 1996). The emergence of concerns about EDCs has resulted in regulations requiring assessment of chemicals for estrogenic activity (Adler et al. 2011; US EPA 1996; US FDA 1996). There are numerous *in vitro* and *in vivo* protocols to identify potential endocrine pathway-mediated effects of chemicals, including interactions with hormone receptors (Jacobs et al. 2008; Rotroff et al. 2013; Shanle and Xu 2011; Sung et al. 2012). However, experimental testing of chemicals is expensive and time-consuming and currently impractical for application to the vast number of synthetic chemicals in use.

Consequently, toxicological data and especially estrogenic activity data are available only for a limited number of compounds (Cohen Hubal et al. 2010; Egeghy et al. 2012; Judson et al. 2009).

The use of *in silico* approaches, such as quantitative structure-activity relationships (QSARs), is an alternative to bridge the lack of knowledge about chemicals when little or no experimental data are available. These structure-based methods are particularly appealing for their ability to predict toxicologically relevant endpoints quickly and at low cost (Muster et al. 2008; Vedani and Smiesko 2009). QSARs have been promoted and their use recognized since the pioneering work of Hansch in the 1960s (Fujita et al. 1964; Hansch et al. 1962; Hansch and Deutsch 1966). The conceptual basis of QSARs is that chemicals with similar structures are hypothesized to exhibit similar behavior in living organisms. Thus, it is possible to predict biological activity of new chemicals based on published experimental data. Several guidance documents to develop these modeling techniques are available in the literature (Dearden et al. 2009; Worth et al. 2005).

Recently, *in vitro* high-throughput screening (HTS) assays have emerged and become a viable tool for large-scale chemical testing (Judson et al. 2011; Kavlock and Dix 2010; Wetmore et al. 2012). HTS generates substantial amounts of data that can be used as a knowledge base to correlate chemical structures to their biological activities. Thus, QSARs can identify key structural characteristics in active chemicals and can use them to virtually screen large chemical libraries. Although there is concern about the overall accuracy of a QSAR model to predict the "true" activity of a particular chemical, accuracy can be high enough to use the results for prioritizing chemicals that are worth subjecting to experimental testing.

With the increasing number of new substances submitted to the U.S. EPA and the European chemicals agency for registration (~1500 chemicals every year), there is a need to prioritize chemicals to speed up the process and lower the overall costs of testing (US EPA 2015). The U.S. Tox21 and EPA's ToxCast projects are screening thousands of chemicals in HTS *in vitro* assays for a broad range of targets (Dix et al. 2007; Judson et al. 2010; Martin et al. 2010). Relevant to this paper, these two projects have in common ~1800 chemicals tested in a battery of 18 ER-related assays (Huang et al. 2014; Judson et al. 2015).

This paper describes the results of CERAPP, a collaborative effort organized by the National Center for Computational Toxicology at the U.S. EPA. The aim of the project was to use ToxCast/Tox21 ER HTS assay data to develop and optimize predictive computational models, and to use their predictions to prioritize a large chemical universe of 32,464 unique chemical structures for further testing. Seventeen research groups from the United States and Europe participated in this project. These groups submitted 40 categorical models and 8 continuous models using different QSAR and structure-based approaches. Most of the newly developed models used a training set consisting of 1,677 chemicals, each assigned a potency score quantifying their ER agonist, antagonist, and binding activities, obtained from a computational network model that integrates data from 18 diverse ER HTS assays (Judson et al. 2015). All models were evaluated and weighted based on their prediction accuracy scores (including sensitivity and specificity) using ToxCast/Tox21 HTS data, as well as an evaluation data set collected from different literature sources. To overcome the limitations of single models, all predictions were combined into a *consensus* model that classified the chemicals into active/inactive binders, agonists, and antagonists and provided estimates of their potency level relative to known reference chemicals.

Materials and methods

Participants and project planning

The 17 international research groups that participated in this project are listed in alphabetic order in Table S1. The goals of the project, outlined in Table S2, were achieved in multiple steps, including chemical structure curation, experimental data preparation from the literature, modeling and prediction, model evaluation, consensus strategy development, and consensus modeling. Each step was assigned to a subgroup of participants according to their interests and areas of expertise.

Data sets

Provided training set. The data that were suggested to be used by the participants as a training set to develop and optimize their models was derived from ToxCast and Tox21 programs (Dix et al. 2007; Huang et al. 2014; Judson et al. 2010). Concentration-response data from a collection of 18 *in vitro* HTS assays exploring multiple sites in the mammalian ER pathway were generated for 1812 chemicals (Judson et al. 2015; US EPA-NCCT 2014b). This chemical library included 45 reference ER agonists and antagonists (including negatives), as well as a wide array of commercial chemicals with known estrogen-like activity (Judson et al. 2015). A mathematical model was developed to integrate the *in vitro* data and calculate an area under the curve (AUC) score, ranging from 0 to 1, which is roughly proportional to the consensus AC50 value across the active assays (Judson et al. 2015). A given chemical was considered active if its agonist or antagonist score was higher than 0.01. In order to reduce the number of potential false positives this threshold can increased to 0.1.

Prediction set. More than 50,000 chemicals (at the level of Chemical Abstracts Service Registry Number [CASRN]) where identified for use in this project as a virtual screening library to be prioritized for further testing and regulatory purposes. This set was intended to include a large fraction of all man-made chemicals to which humans may be exposed. These chemicals were collected from different sources with significant overlap and cover a variety of use classes, including consumer products, food additives, and human and veterinary drugs. The sources include:

(1) Chemicals with documented use and, therefore, with exposure potential (~43,000). Available in the EPA chemical product categories database (CPCat), which is part of the ACToR system (Dionisio et al. 2015; Judson et al. 2008, 2012; US EPA 2014a).

(2) The DSSTox collection of structures (US EPA-NCCT 2014a). A list of ~15,000 curated chemical structures from multiple inventories of environmental interest. In particular, structures for all of the ToxCast and Tox21 chemicals are included.

(3) The Canadian Domestic Substances list (DSL) (Environment Canada 2012). A compiled a list of all substances thought to be in commerce in Canada (~24,000 chemicals). Thus, it includes chemicals with potential human or ecological exposure.

(4) The Endocrine Disruption Screening Program (EDSP) universe of ~10,000 chemicals. EPA's EDSP is required to test certain chemicals for their potential for endocrine disruption (US EPA-NCCT 2014c).

(5) A list of ~15,000 chemicals used as training and test sets for the different models implemented in EPISuite to predict physico-chemical properties (US EPA 2014b).

This virtual chemical library, having undergone stringent chemical structure processing and normalization for use in QSAR modeling study (see chemical curation section here below) and made available for download on the EPA Toxcicity ForeCaster (ToxCast) Data website under CERAPP data (See PredictionSet.zip)(US EPA-NCCT 2016), is intended to be employed for a large number of other QSAR modeling projects, not just those focused on endocrine-related targets.

Experimental evaluation set. A large volume of estrogen-related experimental data has accumulated in the literature over the last two decades. The information on the estrogenic activity of chemicals was mined and curated to serve as a validation set for predictions of the different models. For this purpose, *in vitro* experimental data were collected from different overlapping sources, including EPA's HTS assays, online databases, and other data sets used by participants to train models, namely:

- HTS data from Tox21 project consisting of ~8000 chemicals evaluated in four assays (Attene-Ramos et al. 2013; Collins et al. 2008; Huang et al. 2014; Shukla et al. 2010; Tice et al. 2013), extending beyond the 1,677 used in the training set ;
- The U.S. Food and Drug Administration's Estrogenic Activity Database (EADB), which consists of literature derived ER data for ~8000 chemicals (Shen et al. 2013);
- Estrogenic data for ~2000 chemicals from METI database (METI Ministry of Economy Trade and Industry, Japan 2002); and
- Estrogenic data for ~2000 chemicals from ChEMBL database (Gaulton et al. 2012).

The full data set consisted of more than 60,000 entries, including binding, agonist, and antagonist information for ~15,000 unique chemical structures. For the purpose of this project, this data set was cleaned and made more consistent by removing *in vivo* data, cytotoxicity information, and all ambiguous entries (missing values, undefined/non-standard endpoints, and unclear units). Only 7,547 chemical structures from the experimental evaluation set that overlapped with the CERAPP prediction set, for a total of 44,641 entries, were kept and made available for download on the EPA ToxCast Data website (See EvaluationSet.zip) (US EPA-NCCT 2016). The non-CERAPP chemicals were excluded from the evaluation set (see below). Then, all data entries were categorized into three assay classes: (1) binding, (2) reporter gene / transactivation, or (3) cell proliferation. The training set endpoint to modelis the ER model AUC which parallels the corresponding individual assay AC50 values, and therefore all units for activities in the experimental data set were converted to µM to have approximately equivalent concentration-response values for the evaluation set. Chemicals with cell proliferation assays were considered as actives if they exceeded an arbitrary threshold of 125% proliferation. For entries where testing concentrations were reported in the assay name field, those values were converted to µM and considered as the AC50 value if the compound was reported as active. All inactive compounds were arbitrarily assigned an AC50 value of 1 M.

Chemical structure curation

Chemical structures collected from different public sources contained many duplicates, and inconsistency in the molecular structures. Hence, a structure curation process was carried out to derive a unique set of QSAR-ready structures. All participating groups then used this consistent set of structures for both training and prediction steps. It should be noted that each group likely employed different descriptor calculation software, which could effectively alter structures in some cases. Several different curation approaches were combined into a unique procedure used for this project (Fourches et al. 2010; Wedebye et al. 2013). The free and open-source data-mining environment KNIME was selected to design a curation workflow to process all structures and provide consistent training and prediction sets (Berthold et al. 2007). The workflow performed a series of curation steps, as follows:

(1) The original files containing structures in different formats were parsed, checked for valences, and for the integrity of the required structural information to render the molecules. Invalid entries were corrected by retrieving a new structure from online databases using Web services (Pubchem (NIH 2015), ChemSpider (Royal Society of Chemistry 2015)) or removed if ambiguous.

(2) The first filter was applied to check for the presence of carbon atoms and remove inorganic compounds.

(3) The structures were desalted, and inorganic counter-ions were removed.

(4) The second filter, based on molecular weight, was applied and chemicals exceeding a threshold of 1000 g/mol were removed to speed up molecular descriptor calculations and model calibration.

(5) Valid QSAR modeling practice, requires all chemicals to be structurally consistent by converting tautomers to unique representations. Thus, a series of transformations was applied on the structures to standardize nitro and azide mesomers, keto-enol tautomers, enamine-imine tautomers, ynol-ketene, and other conversions (ChemAxon 2014; Reusch 2013; Sitzmann et al. 2010).

(6) These transformations were followed by neutralizing the charged structures, when possible, and removing the stereochemistry information.

(7) Explicit hydrogen atoms were added, and structures were aromatized according to Hückel's rules implemented in KNIME (Berthold et al. 2007).

(8) The duplicates were removed using InChI (IUPAC International Chemical Identifier) codes, because these are unequivocal identifiers.

(9) The final filter was applied to remove chemicals containing metals which often cause problems in molecular descriptor calculations.

Both training and prediction sets were processed by the same structure curation workflow. At the end of this procedure, 32,464 unique structures (hereafter referred to as the 32K set) remained in the prediction set and 1,677 in the training set. These two data sets are made available for download in SDF format on the EPA ToxCast Data website (See TrainingSet,zip and PredictionSet.zip) (US EPA-NCCT 2016). The identity of these chemicals (name, CASRN) was not provided to the participating modeling groups during the modeling process.

Modeling approaches

The participant groups adopted different approaches and used several software programs (proprietary or open-source [commercial or free]) to calibrate categorical and continuous models to the training data (Table 1). A categorical model is one that provides an active/inactive call for each chemical, whereas a continuous model provides a prediction of the potency (in µM) for each active chemical. Models were developed using both well-known and innovative methods including partial least-squares (PLS) (Ståhle and Wold 1987; Wold et al. 2001), partial least-squares discriminant analysis (PLS-DA) (Frank and Friedman 1993; Nouwen et al. 1997), decision forest (DF) (Hong et al. 2005, 2004; Tong et al. 2003; Xie et al. 2005), three-dimensional quantitative spectral data-activity relationship (3D-QSDAR) (Beger et al. 2001; Beger and Wilkes 2001; Slavov et al. 2013), support vector machines (SVM) (Cristianini and Shawe-Taylor 2000), *k* nearest neighbors (kNN) (Cover and Hart 1967; Kowalski and Bender

1972), associative artificial neural networks (ASNN) (Tetko 2002a, 2002b), PASS algorithm derived from Naïve Bayes classifier (Poroikov et al. 2000), self-consistent regression with radial basis function interpolation (RBF-SCR) (Zakharov et al. 2014), OCHEM machine learning methods (Tetko et al. 2014), docking and *consensus* of different approaches (Horvath et al. 2014; Ng et al. 2014; Sushko et al. 2011). The set of 1677 chemicals provided by EPA was used by more than 90% of the participating groups as a training set to fit their models (Judson et al. 2015), but some preexisting models were also used, that had been trained using other data sets from the literature such as METI (METI Ministry of Economy Trade and Industry, Japan 2002). In addition, each group performed its own analysis to select the appropriate chemicals to be considered as a training set according to their particular modeling procedure. For descriptor calculation and docking procedures, some of the programs used were LeadScope (Roberts et al. 2000), PADEL (Yap 2011), Qikprop (Schrödinger, LLC 2011), multilevel and quantitative neighborhoods of atoms (MNA, QNA) used by GUSAR and PASS (Filimonov et al. 2009; Poroikov et al. 2000), DRAGON (Talete srl 2012), Mold2 (Hong et al. 2008, 2012), GLIDE (Schrödinger) (Schrödinger, LLC 2011), AutoDock (Goodsell et al. 1996), ISIDA (Varnek et al. 2008) and other fingerprint generators. Some of the participants applied feature selection techniques, such as genetic algorithms (GAs) (Davi 1991) and random forest (RF) (Breiman 2001). These techniques were applied after calculating descriptors to reduce collinearity and variable dimensionality to keep only the most informative descriptors in the models.

Evaluation procedure for the categorical and continuous models

All molecular structures of chemicals collected for the evaluation set from the different sources were curated and standardized using the previously described KNIME workflow (See step 2 in

Table S2). All data used as the evaluation set for categorical and continuous models are available on the EPA ToxCast website (See EvaluationSet.zip) (US EPA-NCCT 2016).

Standard InChI codes were generated in KNIME and used to identify the chemicals. Data-mining tools available in the KNIME environment were used to concatenate and unify the different information fields from the different sources (CASRN, chemical name, original structure, standardized structure, InChI code, assay name, assay class, protein subtype, species, endpoint name, endpoint value, endpoint unit, and literature reference). Even though ToxCast chemicals were used in the training sets of many models, they were not removed from the evaluation set to investigate how the predictions will perform on the literature data knowing that there are differences between the AUC values and the literature data. Also because the sources from which the evaluation set was collected are not fully verified (we cannot assume that all cytotoxicity information was already fully cleaned).,

Evaluation set for categorical models. An important issue with the literature-derived evaluation set (discussed further below) was the inconsistency of the results from different sources. To minimize this, the available entries for each chemical structure were grouped into binders, agonists, and antagonists. The results were then categorized into active and inactive classes using all available literature sources by applying three rules.

(1) If, for a specific chemical within one of the three classes (binding, agonist and antagonist each apart), the disagreement among the different sources exceeds 20% (e.g. 2 sources indicating active agonist and 3 indicating inactive agonist), that chemical was removed from the evaluation data set of that specific class.

(2) If a chemical was an active agonist or antagonist, it also was considered as an active binder if the information was not available.

(3) If a chemical was an inactive agonist and inactive antagonist, it was considered also as non-binder if the information was not available.

This procedure resulted in a total of 7,522 unique chemical structures with activity data to be used for evaluation of the categorical models (See Table 2 and available for download on the EPA ToxCast website, EvaluationSet.zip) (US EPA-NCCT 2016).

Evaluation set for continuous models. For active chemicals with available quantitative information from concentration-response assays, the $\log_{10}$-median of the literature values was calculated. Only entries with equivalent endpoints were considered (e.g. PC50 and EC50). This resulted in 7,253 unique chemicals with quantitative information (See Table 3 and available for download on the EPA ToxCast website, see EvaluationSet.zip) (US EPA-NCCT 2016).. To reduce the variability that increased with the disparate literature sources, the chemicals with quantitative information were categorized into five potency activity classes: inactive, very weak, weak, moderate, and strong. These five classes were used to evaluate the quantitative predictions. A list of 36 known active and inactive reference chemicals was used for calibrating the mapping from quantitative potency values to the activity potency classes (Judson et al. 2015). These same chemicals were used to validate the mathematical model used to generate the AUC values for the training set. The following thresholds were applied to the concentration-response values.

(1) Strong: Activity concentration below 0.09 μM

(2) Moderate: Activity concentration between 0.09 and 0.18 μM

(3) Weak: Activity concentration between 0.18 and 20 μM

(4) Very Weak: Activity concentration between 20 and 800 μM

(5) Inactive: Activity concentration higher than 800 μM

The five classes were assigned scores from 0 (inactive) to 1 (strong) with 0.25 increments. Then, for each chemical, the arithmetic mean of the scores of the merged entries from different literature sources was calculated. A new class was assigned to the merged entries according to the following thresholds.

(1) Strong: Average score > 0.75

(2) Moderate: 0.5 < Average score between <= 0.75

(3) Weak: 0.25 < Average score <= 0.5

(4) Very weak: 0 < Average score <= 0.25

(5) Inactive: Average score = 0

The number of entries in each class for binding, agonist, and antagonist are summarized in Table 3.

Evaluation procedure. This section is focusing on the categorical models for their high number compared to the continuous models. The procedure used to evaluate the predictions of the participant groups was based on the categorical and continuous experimental data from ToxCast and the evaluation set from the literature. All continuous and categorical models for binding, agonist, and antagonist were evaluated separately on the overlap between their predicted chemicals and the following sets of chemicals (See Table S3).

(1) Chemicals in EPA's ToxCast dataset (n= 1,529 chemicals after excluding those in the ambiguous AUC range of 0.01 to 0.1).

(2) All chemicals in the full literature data (all literature sources combined).

(3) All chemicals with at least two literature sources

(4) All chemicals from the literature data excluding the very weak actives

(5) Chemicals within the applicability domain (AD) of each model (if provided)

(6) Chemicals remaining after applying the previous 3 filters in steps 3, 4 and 5 to reduce ambiguous predictions (single literature source, very weak actives, and predictions outside the AD)

To evaluate the models on different criteria, we first determined the sensitivity (fraction of accurately predicted actives out of all actives), specificity (fraction of accurately predicted inactives out of all inactives), and balanced accuracy (average of sensitivity and specificity) for each subgroup of chemicals according to each model. We then used BA values to derive two summary scores for each model, as described below.

Score_1. Evaluation includes BA of each of the six steps weighted by the fraction of predicted chemicals of the same step as well as the fraction of the predicted chemicals out of the full prediction set. This score favors models with a wider AD and those predicting a maximum number of chemicals.

$$score\_1 \;\; = {}^{1}/_{3}\left(\frac{BA_{ToxCast} * N\_pred_{ToxCast}}{N_{ToxCast}} + \frac{N\_pred}{N\_total} + {}^{1}/_{N_{filters}} \sum_{i=1}^{N_{filters}} \frac{BA_i * N\_pred_i}{N\_total_i}\right) \quad [1]$$

where BA is balanced accuracy, N_pred is the number of predicted chemicals by a specific model, N_total is the total number of chemicals in the prediction set, $N_{filters}$ represents the number of 5 filters applied to the evaluation set chemicals and *i* the steps 2, 3, 4, 5 and 6.

Score_2. Evaluation includes the BA of the model on the ToxCast data, and the BA on the unambiguous chemicals: , i.e., the subgroup of chemicals from the literature that remained after excluding chemicals with only 1 literature source, very weak chemicals, and chemicals outside of the AD, if provided. It favors models that focused on predicting more accurately but, potentially, with a narrower AD.

$$score\_2 \quad = {}^{1}\!/_{2}\left(BA_{ToxCast} + BA_{all\ filters}\right) \ [2]$$

The quantitative predictions were evaluated as categorical models (using the BA) of the five classes after converting the numerical predictions to potency classes as defined earlier. Scores of the continuous models were calculated using equation (2).

Consensus modeling

The *consensus* predictions were generated for binders, agonists, and antagonists separately. For each chemical we derived the average Score 2 value for all categorical models that predicted the chemical as active, and the average Score 2 value for all categorical models that predicted the chemical as inactive, and used the higher of the two averages to classify the chemical as active or inactive. Models that did not provide a prediction for the chemical in question were not included when deriving the average scores. We used Score 2 to derive the consensus classifications because its value for individual models is not penalized for the number of chemicals not predicted by the model. Also, the concordance among models on both active and inactive classes was calculated for each chemical as the fraction of models with positive and negative prediction, respectively.

Considering only the models that provided predictions, the sum of the concordance among models for actives and inactives is equal to 1. Because most models were associated with comparable scores, the average score used to classify chemicals was mostly in agreement with model concordance; i.e., the average score for actives is high when the concordance among the models with active predictions is high and vice versa. The few exceptions were noticed when model concordance was around 0.5, which means only one or two models were driving the classification.

For continuous predictions, the weight (w) for each chemical i was calculated from the scores as follows:

$$w_i = score_i / \sum_{j=1}^{n} score_j \quad [3]$$

where n is the total number of models that provided predictions for the chemical i, and $score_j$ is the score of the j^{th} model predicting chemical i.

Next, the *consensus* potency level C_i of each chemical was determined using the predicted potency classes P_j of the n available models and their corresponding weights w as follows:

$$C_i = \sum_{j=1}^{n} w_j \cdot P_j \quad [4]$$

Results and discussion

Models and evaluation

A total of 48 models were received from the 17 participant groups. Each group provided at least one categorical model for binding. Only 8 groups built models for agonists, and 6 groups built models for antagonists. The limited number of models for agonists and antagonists was the result of the low number of actives, which caused the training set to be highly unbalanced. The total number of models in each class (Table 1, Table S3 and Table S5) was:

(1) binding models: 21 categorical and 3 continuous,

(2) agonist models: 11 categorical and 3 continuous, and

(3) antagonist models: 8 categorical and 2 continuous.

The participating groups provided predictions for uneven fractions of the 32k set. AD information on model predictions was provided by only six groups. All predictions for the individual models are provided on the EPA ToxCast website (See Models.zip) (US EPA-NCCT 2016).

The same evaluation procedure was applied to all models following the previously described steps. Note that some models were built using training sets other than what was provided in CERAPP and that these alternative training sets were not all publicly available. Hence, none of the training set chemicals were excluded from the evaluation sets (Table 1). Each model was evaluated on the overlap between the predicted chemicals and the two previously mentioned data sets: (1) ToxCast data and (2) the evaluation set collected from the literature. The evaluation results for categorical models are summarized in Table S3. The detailed statistics, including sensitivity and specificity, are provided in Table S4.

Most compounds were predicted as inactives and the models seemed to be more in agreement in predicting inactives than active compounds. Only 757 chemicals (2.33%) are predicted as actives by more than 75% of binding models. The agreement among the binding models for the 32k set of the prediction set is illustrated in Figure S1.

Most categorical models (binding, agonist, and antagonist) are associated with high balanced accuracies on the ToxCast data (> 0.8), with no clear difference between models that used it as a training set and those that did not (See Table S3). However, for the evaluation set from the literature, the BA is clearly lower for all models (<0.7). Nonetheless, the BA increased after removing chemicals with only one source from the literature data. This result could mean that this first filter (i.e., removing chemicals with limited information in the literature for being either positive or negative) reduced the uncertainty in the experimental data from the literature.

This is in agreement with related studies showing that the results of QSAR models may change depending on the robustness of the experimental values (Steinmetz et al. 2014). The second filter (i.e., removing very weak actives) also increased the BA, which suggests that the literature data may contain a number of false positives. Alternatively, the *in vitro* assays used by ToxCast/Tox21 only test chemicals up to 100 µM, so very weak chemicals may not be picked up by these assays and some of the literature reports may have tested chemicals up to much higher concentrations.

Finally, removing predictions outside the AD did not show improvement of the BA of the categorical models (See Table S3). This is in agreement with literature sources showing that predictions outside the AD are not always less accurate than those within its limits (Sahigara et al. 2012). The performance of most models showed a clear improvement of 0.05 to 0.1 on the BA after applying all the filters on the literature data to keep only the unambiguous chemicals. We believe that this effectively reduced the uncertainty of the literature sources. This step also highlighted differences between ToxCast and the literature data and confirmed the existence of uncertainty in the literature data. Uncertainty and data discordance was also reported in literature review of *in vivo* uterotrophic bioassays (Kleinstreuer et al. 2015).

The calculated scores for categorical models (Table S3) take into consideration the whole prediction set (Score_1) and the accuracy of the model on its most reliable predictions (Score_2). The models that provided predictions for the whole or most of the 32k set of chemicals, and had wide ADs, showed high Score_1 values (Umeå 0.82, OCHEM 0.83). Whereas models with predictions for smaller fractions of the prediction set and narrow AD showed better Score_2 values (UNIMIB_2 0.85, UNIBA 0.80). NIH_NCI_GUSAR (0.87 and 0.84) and

FDA_NCTR_DBB (0.88 and 0.84) showed the highest values for both Score_1 and Score 2. Part of the differences among model scores could result from the uncertainty in the literature data.

The BAs of all antagonist models was low compared with binding and agonist models (Table S3). This may be due to the highly unbalanced training set with a low number of active antagonist chemicals. Additionally, antagonism activity (in either ToxCast or the literature) can be confounded with cytotoxicity because antagonist transactivation assays are loss-of-signal assays.

The predictions of all continuous models were first converted to five classes using the list of reference chemicals as described in the evaluation set section (materials and methods). The predictions were then evaluated on the ToxCast data and the literature data to calculate the average of BA of the different evaluation steps as the score of each model (See Table S5). All models showed high BA on ToxCast data and relatively good BA on the evaluation set.

***Consensus* model**

The *consensus* predictions were first evaluated on the ToxCast data and then on the evaluation set from the literature. The total number of predicted active binders was 2661 out of the 32k set of chemicals (8.2%) based on the method described in the Materials and methods section *Consensus* modeling.

Confusion matrices (Table 4) and prediction statistics (Table 5) revealed a clear accuracy difference between the categorical *consensus* for binding on the ToxCast data and on the evaluation set. This difference could result from the fact that the ToxCast data, based on a model with inputs from 18 different assays, were used by most of the models as a training set, which we presume reduces the uncertainty. This is in contrast to the literature data, where the number of sources per chemical varied from one to a few hundreds. When only the subset of the evaluation set with more than six literature sources per chemical was considered, a large increase in the sensitivity was noticed (0.23 to 0.85).

To better understand the effect of the number of sources on the classification accuracy, ROC plots were made using the fraction of the binding models in each class as a threshold for the classification predictions and increasing the number of literature sources of the evaluation set. The ROC plot shows an improvement of the classification accuracy of the *consensus* model as the number of sources increases (Figure 1). Note that the same level of consistency (i.e., 80%) was required to merge the sources regardless of the number of sources (See rule 1 in Section Evaluation set for categorical models). This could lead to the conclusion that the low classification accuracy on the full literature data is not because of a lack of accuracy of the *consensus* predictions, but rather to noise and experimental uncertainty in the literature data. We assume that the high number of false negatives in the confusion matrix of Table 4 is caused by false positives in the full literature data for chemicals tested only a small number of times. Thus, by considering a higher number of sources (i:e:, 6), the number of false positives is reduced from the evaluation set and so the number of predicted false negatives decreased. This is in agreement with what was observed in the literature (Steinmetz et al. 2014).

Corrections to the *consensus* model

The first step of *consensus* modeling was conducted in an independent way for the categorical and continuous models on binding, agonist, and antagonist predictions. This led to a number of inconsistencies because some chemicals were predicted as active in categorical predictions but inactive in quantitative and vice versa. In addition, some chemicals were predicted as active agonists or antagonists but non-binders. To make all predictions more consistent, a number of corrections were applied on the first *consensus* predictions. Because the goal of this project was to help in a regulatory prioritization procedure, the modifications aimed to reduce the number of false negatives but without adding an excess of false positives. The rules that were followed to obtain the final *consensus* predictions are as follows:

(1) If a chemical *i* is active in the categorical *consensus*, then it is considered active also in the quantitative *consensus*.

(2) If a chemical *i* is active in the quantitative *consensus* and predicted as active by at least three categorical models, then it is considered active also in the categorical consensus.

(3) If a chemical *i* is predicted active by less than three categorical models, then it is considered inactive also in quantitative *consensus*.

These 3 rules were applied on the agonist and antagonist *consensus* models first, then on the binding consensus. A fourth rule was added to establish consistency between agonist and antagonist *consensus* models and the binding *consensus* model.

(4) If a chemical *i* is an active agonist or active antagonist, then it is considered as active in categorical binding *consensus,* and its potency level in the quantitative binding *consensus* is made equal to its potency level as agonist/antagonist.

An analysis of variance in concordance in each potency level of the active chemicals in the continuous models (very weak, weak, moderate, and strong) is presented as a box-plot in Figure 2. Based on this figure we noticed a correlation between the concordance of the categorical models and the potency level of active chemicals. This implies that models are more in agreement for strong actives and that the weaker a chemical is the more difficult it is to accurately predict. Therefore, the very weak chemicals are the main source of discordance among the different *in silico* models and also are the most uncertain experimentally. This relationship between positive concordance (agreement between models on predictions for active chemicals) and potency level for active chemicals can be used to set a quantitative prediction to the newly reclassified active chemicals using the previously mentioned rule 1 of the corrections applied to the consensus predictions. The following thresholds were considered for each potency level.

(1) Strong: Concordance among models >= 0.9

(2) Moderate: 0.75 <= Concordance among models < 0.9

(3) Weak: 0.6 <= Concordance among models < 0.75

(4) Very weak: Concordance among models < 0.6

.

After applying the four correction rules on consensus predictions, the total number of chemicals predicted as actives increased from 2661 to 4001, which corresponds to 12.3% of the total number of the prediction set (32,464). Table 6 shows the number of reclassified chemicals based on each one of the 4 correction rules applied to the consensus predictions. After this step, the predicted activity of several chemicals has changed. The structural information of chemicals and the predictions of the *consensus* model for the whole 32k set are provided on the EPA ToxCast website (See PredictionSet.zip) (US EPA-NCCT 2016).

The confusion matrices and statistics for the binding categorical *consensus* model after modifications evaluated on ToxCast data and the literature data are presented in Table 7 and Table 8, respectively. The effect of the number of sources on the classification accuracy of the *consensus* model is illustrated by a bar plot in Figure S2. This figure shows an improvement of sensitivity with the increase in the number of literature sources in the evaluation set (from ~0.3 with at least 1 source to >0.6 with 6 sources and more). This is translated into an increase in BA, whereas specificity is almost constant (~0.9) because of the high number of inactives compared to active compounds.

The results of this project and the ToxCast data used as the training set, are published online in the EDSP21 dashboard, together with other structural and experimental assay information (See “Consensus CERAPP QSAR ER Model Predictions” under “Chemical Summary” tab on http://actor.epa.gov/edsp21) (US EPA-NCCT 2014b). A comparison of the single classification models to the *consensus* predictions for the whole 32k set of chemicals is provided in Table S6. The calculations are done using the categorical consensus predictions as the “observed response”.

For regulatory or prioritization purposes, one could use a looser definition of active (.e. allow more disagreement among models) in order to further reduce the chance of false negatives. Figure 3 shows the number of chemicals that can be predicted as potential actives by the categorical consensus for binding using various positive concordance (agreement on actives between the included models) thresholds. When this threshold is set to 0.2, a total of additional

6,742 more chemicals can be added to the potential positives (this refers to the available binding models). This figure also shows the BA variations at different number of literature sources in the literature. Balanced accuracy increases as the concordance threshold increases from 0 to 0.2 because sensitivity increases (false negatives decrease) as the number of chemicals classified as active increases. For chemicals with the highest data quality (seven or more sources), the BA curve reaches a plateau at concordance thresholds of 0.4–0.5, and the number of chemicals classified as active is consistent with the number of active chemicals predicted from our consensus model (n = 4001.) However, higher concordance thresholds result in declining BA due to increasing numbers of false positive predictions (i.e., decreasing specificity).

Conclusion

The collaborative efforts of CERAPP participants resulted in *consensus* predictions of the ability of chemicals to interact with ER. Up to 48 separately developed categorical and continuous models were received from 17 research groups from the United States and Europe. Separate models were built for agonist, antagonist, and binding activity. The models were applied to a large collection of 32,464 chemical structures that approximate the human exposure universe (chemicals with potential human exposure). A KNIME workflow was developed to carefully curate the large collection of chemical structures to ensure consistency in model development and evaluation. Most of the models were trained using activities derived from a dataset combining 18 *in vitro* assays from ToxCast probing various points of the ER pathway. Models, then, were evaluated using the ToxCast data plus a collection of ER *in vitro* data from the literature. Categorical predictions were after that combined into a consensus to classify the chemicals into actives and inactives, while continuous predictions were combined to classify the actives into 4 different potency classes: very weak, weak, moderate, and strong.

One major observation was that most models had comparable performances, independent of the methods used, with a slight improvement for models with narrow ADs. A second and, perhaps, more important observation is that the most concordant predictions come from comparing the *consensus* of many models with a *consensus* of many literature sources. For instance, when comparing the *consensus* of the categorical binding models with the evaluation set from the literature for chemicals with seven or more sources, we achieve a balanced accuracy of about 90% (Table 8).

We propose several important conclusions from our results. First, there does not appear to be an optimal modeling approach (combination of descriptor set, feature selection, or machine learning algorithm) that will solve the QSAR/docking problem and achieve perfect prediction accuracies. Second, there are inherent limitations to the accuracy of the data being used to train QSAR and docking models. Our analysis of the literature data showed a disagreement in the reported activity of many chemicals. The sources of discrepancy include limits to the concentration ranges tested, true differential activity among tissue sources (e.g., the presence of selective ER modulators, SERMs), and a variety of experimental artifacts and errors. Figure 2 shows that the most consistent predictions are achieved for the most potent compounds, whereas weaker compounds are called inactive by some laboratories because these compounds were not tested at a high enough concentration. So chemicals with very weak activity would be more likely to be incorrectly classified as inactive than more potent chemicals. Therefore, 100% accuracy cannot be achieved due to these limitations in the experimental data used for training and evaluation. Figures 1 and 3 help to illustrate this point by showing that higher consistency in the experimental data is associated with an increase in the concordance among model predictions. But this comes at the cost of excluding parts of the experimental data. So, just as every model has limitations, every *in vitro* assay also has inherent variability in its results.

The major purpose of this study was to identify potential ER actives out of the large universe of chemicals to which humans potentially are exposed using a *consensus* of *in silico* models to overcome the limitations of single models. Most of the chemicals in this collection were predicted to be negatives, with a high agreement among the individual models. The disagreement was the highest for chemicals with weak activity (Figure 2). This disagreement is driven by the difficulties in experimentally assessing the activity of these weak chemicals. In total, the consensus predicted 4001 chemicals as actives. The testing of these active chemicals will be prioritized from the most potent to the least according to the continuous model *consensus* predictions. There are 6,742 more chemicals that 20% to 50% of the models predicted to be positive, which could also be candidates for follow-up. Although this large number of chemicals (~10,000 in total) appears to be a daunting set to evaluate experimentally, this is equivalent in size to the current Tox21 library already being tested for activity in ER and many other targets.

In summary, this project demonstrates the feasibility of screening a large and toxicologically relevant library of chemical structures in an extensive battery of QSAR and docking models to meet important goals in human and environmental health. ER provides a good initial case because of the ready availability of experimental data and preexisting models. However, through the ToxCast and Tox21 programs, and through other large scale data-integration projects, equivalently large data sets will become available for multiple other targets of environmental importance.

References

Adler S, Basketter D, Creton S, Pelkonen O, Benthem J van, Zuang V, et al. 2011. Alternative (non-animal) methods for cosmetics testing: current status and future prospects—2010. Arch. Toxicol. 85:367–485.

Attene-Ramos MS, Miller N, Huang R, Michael S, Itkin M, Kavlock RJ, et al. 2013. The Tox21 robotic platform for the assessment of environmental chemicals--from vision to reality. Drug Discov. Today 18:716–723; doi:10.1016/j.drudis.2013.05.015.

Beger RD, Buzatu DA, Wilkes JG, Lay JO, Jr. 2001. 13C NMR quantitative spectrometric data–activity relationship (QSDAR) models of steroids binding the aromatase enzyme. J Chem Inf Comput Sci 41: 1360– 1366.

Beger RD, Wilkes JG. 2001. Developing 13C NMR quantitative spectrometric data–activity relationship (QSDAR) models of steroid binding to the corticosteroid binding globulin. J Comput Aid Mol Des 15: 659– 669.

Berthold MR, Cebron N, Dill F, Gabriel TR, Kötter T, Meinl T, et al. 2007. KNIME: The Konstanz Information Miner. Stud. Classif. Data Anal. Knowl. Organ. GfKL 2007.

Birnbaum LS, Fenton SE. 2003. Cancer and developmental exposure to endocrine disruptors. Environ. Health Perspect. 111: 389–394.

Breiman L. 2001. Random forests. Mach. Learn. 45:5–32.

ChemAxon. 2014. ChemAxon Standardizer–Cheminformatics platforms and desktop applications. Available: http://www.chemaxon.com/products/standardizer/ [accessed 26 November 2014].

Cohen Hubal EA, Richard A, Aylward L, Edwards S, Gallagher J, Goldsmith M-R, et al. 2010. Advancing Exposure Characterization for Chemical Evaluation and Risk Assessment. J. Toxicol. Environ. Health Part B 13:299–313.

Colborn T, vom Saal FS, Soto AM. 1993. Developmental effects of endocrine-disrupting chemicals in wildlife and humans. Environ. Health Perspect. 101: 378–384.

Collins FS, Gray GM, Bucher JR. 2008. Toxicology. Transforming environmental health protection. Science 319:906–907.

Cover T, Hart P. 1967. Nearest neighbor pattern classification. IEEE Trans. Inf. Theory 13:21–27.

Cristianini N, Shawe-Taylor J. 2000. *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods*. 1st ed. Cambridge University Press.

Davi L. 1991. *Handbook of Genetic Algorithm*. Van Nostrand Reinhold, New York.

Davis DL, Bradlow HL, Wolff M, Woodruff T, Hoel DG, Anton-Culver H. 1993. Medical hypothesis: Xenoestrogens as preventable causes of breast cancer. Environ. Health Perspect. 101: 372–377.

Dearden JC, Cronin MTD, Kaiser KLE. 2009. How not to develop a quantitative structure–activity or structure–property relationship (QSAR/QSPR). SAR QSAR Environ. Res. 20:241–266.

Diamanti-Kandarakis E, Bourguignon J-P, Giudice LC, Hauser R, Prins GS, Soto AM, et al. 2009. Endocrine-Disrupting Chemicals: An Endocrine Society Scientific Statement. Endocr. Rev. 30:293–342.

Dionisio KL, Frame AM, Goldsmith MR, Wambaugh JF, Liddell A, Cathey T, et al. 2015. Exploring Consumer Exposure Pathways and Patterns of Use for Chemicals in the Environment. Toxicol. Rep; doi:10.1016/j.toxrep.2014.12.009 [Online 2 January 2015].

Dix DJ, Houck KA, Martin MT, Richard AM, Setzer RW, Kavlock RJ. 2007. The ToxCast program for prioritizing toxicity testing of environmental chemicals. Toxicol. Sci. Off. J. Soc. Toxicol. 95:5–12.

Egeghy PP, Judson R, Gangwal S, Mosher S, Smith D, Vail J, et al. 2012. The exposure data landscape for manufactured chemicals. Sci. Total Environ. 414:159–166.

Environment Canada. 2012. DSL (Domestic Substances List). Available: http://www.ec.gc.ca/lcpe-cepa/default.asp?lang=En&n=5F213FA8-1&wsdoc=D031CB30-B31B-D54C-0E46-37E32D526A1F [accessed 4 November 2012].

Filimonov DA, Zakharov AV, Lagunin AA, Poroikov VV. 2009. QNA-based “Star Track” QSAR approach. SAR QSAR Environ. Res. 20:679–709.

Fourches D, Muratov E, Tropsha A. 2010. Trust, but verify: On the importance of chemical structure curation in cheminformatics and QSAR modeling research. J. Chem. Inf. Model. 50: 1189–1204.

Frank IE, Friedman JH. 1993. A statistical view of some chemometrics regression tools. Technometrics 35: 109– 135.

Fujita T, Iwasa J, Hansch C. 1964. A new substituent constant, p, derived from partition coefficients. J Am Chem Soc 86: 5175– 5180.

Gaulton A, Bellis LJ, Bento AP, Chambers J, Davies M, Hersey A, et al. 2012. ChEMBL: a large-scale bioactivity database for drug discovery. Nucleic Acids Res. 40:D1100–D1107.

Goodsell DS, Morris GM, Olson AJ. 1996. Automated docking of flexible ligands: applications of AutoDock. J. Mol. Recognit. JMR 9:1–5.

Hansch C, Deutsch EW. 1966. The structure-activity relationship in amides inhibiting photosynthesis. Bibl. Laeger 112: 381–391.

Hansch C, Maloney PP, Fujita T, Muir RM. 1962. Correlation of biological activity of phenoxyacetic acids with Hammett substituent constants and partition coefficients. Nature 194: 178–180.

HILEMAN B. 1994. Environmental Estrogens linked to Reproductive Abnormalities, Cancer. Chem. Eng. News Arch. 72:19–23.

Hong H, Slavov S, Ge W, Qian F, Su Z, Fang H, et al. 2012. Mold2 Molecular Descriptors for QSAR. In *Statistical Modelling of Molecular Descriptors in QSAR/QSPR* (M. Dehmer, K. Varmuza, and D. Boncheveds.), pp. 65–109, Wiley-VCH Verlag GmbH & Co. KGaA.

Hong H, Tong W, Perkins R, Fang H, Xie Q, Shi L. 2004. Multiclass Decision Forest--a novel pattern recognition method for multiclass classification in microarray data analysis. DNA Cell Biol. 23:685–694.

Hong H, Tong W, Xie Q, Fang H, Perkins R. 2005. An in silico ensemble method for lead discovery: decision forest. SAR QSAR Environ. Res. 16:339–347.

Hong H, Xie Q, Ge W, Qian F, Fang H, Shi L, et al. 2008. Mold(2), molecular descriptors from 2D structures for chemoinformatics and toxicoinformatics. J. Chem. Inf. Model. 48:1337–1344.

Horvath D, Brown JB, Marcou G, Varnek A. 2014. An Evolutionary Optimizer of libsvm Models. Challenges 5:450–472.

Huang R, Sakamuru S, Martin MT, Reif DM, Judson RS, Houck KA, et al. 2014. Profiling of the Tox21 10K compound library for agonists and antagonists of the estrogen receptor alpha signaling pathway. Sci. Rep. 4.

Jacobs M, Janssens W, Bernauer U, Brandon E, Coecke S, Combes R, et al. 2008. The Use of Metabolising Systems for In Vitro Testing of Endocrine Disruptors. Curr. Drug Metab. 9:796–826.

Judson R, Richard A, Dix D, Houck K, Elloumi F, Martin M, et al. 2008. ACToR--Aggregated Computational Toxicology Resource. Toxicol. Appl. Pharmacol. 233:7–13.

Judson R, Richard A, Dix DJ, Houck K, Martin M, Kavlock R, et al. 2009. The Toxicity Data Landscape for Environmental Chemicals. Environ. Health Perspect. 117:685–695.

Judson RS, Houck KA, Kavlock RJ, Knudsen TB, Martin MT, Mortensen HM, et al. 2010. In vitro screening of environmental chemicals for targeted testing prioritization: the ToxCast project. Environ. Health Perspect. 118:485–492.

Judson RS, Kavlock RJ, Setzer RW, Hubal EAC, Martin MT, Knudsen TB, et al. 2011. Estimating toxicity-related biological pathway altering doses for high-throughput chemical risk assessment. Chem. Res. Toxicol. 24:451–462.

Judson RS, Magpantay FM, Chickarmane V, Haskell C, Tania N, Taylor J, et al. 2015. Integrated Model of Chemical Perturbations of a Biological Pathway Using 18 In Vitro High Throughput Screening Assays for the Estrogen Receptor. Toxicol. Sci. kfv168; doi:10.1093/toxsci/kfv168.

Judson RS, Martin MT, Egeghy P, Gangwal S, Reif DM, Kothiya P, et al. 2012. Aggregating Data for Computational Toxicology Applications: The U.S. Environmental Protection Agency (EPA) Aggregated Computational Toxicology Resource (ACToR) System. Int. J. Mol. Sci. 13:1805–1831.

Kavlock R, Dix D. 2010. Computational toxicology as implemented by the U.S. EPA: providing high throughput decision support tools for screening and assessing chemical exposure, hazard and risk. J. Toxicol. Environ. Health B Crit. Rev. 13:197–217.

Kavlock RJ, Daston GP, DeRosa C, Fenner-Crisp P, Gray LE, Kaattari S, et al. 1996. Research needs for the risk assessment of health and environmental effects of endocrine disrupters: A report of the U.S. EPA-sponsored workshop. Environ. Health Perspect. 104: 715–740.

Kleinstreuer NC, Ceger PC, Allen DG, Strickland J, Chang X, Hamm JT, et al. 2015. A Curated Database of Rodent Uterotrophic Bioactivity. Environ. Health Perspect.; doi:10.1289/ehp.1510183.

Kowalski BR, Bender CF. 1972. The K-nearest neighbor classification rule (pattern recognition) applied to nuclear magnetic resonance spectral interpretation. Anal. Chem. 44: 1405–1411.

Kuiper GG, Carlsson B, Grandien K, Enmark E, Häggblad J, Nilsson S, et al. 1997. Comparison of the ligand binding specificity and transcript tissue distribution of estrogen receptors alpha and beta. Endocrinology 138:863–870.

Mahoney MM, Padmanabhan V. 2010. Developmental programming: Impact of fetal exposure to endocrine disrupting chemicals on gonadotropin-releasing hormone and estrogen receptor mRNA in sheep hypothalamus. Toxicol. Appl. Pharmacol. 247:98–104.

Martin MT, Dix DJ, Judson RS, Kavlock RJ, Reif DM, Richard AM, et al. 2010. Impact of environmental chemicals on key transcription regulators and correlation to toxicity end points within EPA's ToxCast program. Chem. Res. Toxicol. 23:578–590.

METI Ministry of Economy Trade and Industry, Japan. 2002. Current Status of Testing Methods Development for Endocrine Disrupters. 6th Meeting of the Task Force on Endocrine Disrupters Testing and Assessment (EDTA). Tokyo, Japan. http://www.meti.go.jp/english/report/data/gEndoctexte.pdf

Mueller SO, Korach KS. 2001. Estrogen receptors and endocrine diseases: lessons from estrogen receptor knockout mice. Curr. Opin. Pharmacol. 1:613–619.

Muster W, Breidenbach A, Fischer H, Kirchner S, Müller L, Pähler A. 2008. Computational toxicology in drug development. Drug Discov. Today 13:303–310.

Ng HW, Zhang W, Shu M, Luo H, Ge W, Perkins R, et al. 2014. Competitive molecular docking approach for predicting estrogen receptor subtype α agonists and antagonists. BMC Bioinformatics 15:S4.

NIH. 2015. The PubChem Project. Available: http://pubchem.ncbi.nlm.nih.gov/ [accessed 26 January 2015].

Nouwen J, Lindgren F, Karcher W. 1997. Classification of environmentally occurring chemicals using structural fragments and PLS discriminant analysis. Env. Sci Technol 31: 2313–2318.

OCHEM. 2015. CERAPP models. Available: https://ochem.eu/article/71005 [accessed 12 January 2015].

Poroikov VV, Filimonov DA, Borodina YV, Lagunin AA, Kos A. 2000. Robustness of biological activity spectra predicting by computer program PASS for noncongeneric sets of chemical compounds. J Chem Inf Comput Sci 40: 1349– 1355.

Reusch W. 2013. Examples of chemical reactions. Available: http://www2.chemistry.msu.edu/faculty/reusch/virttxtjml/react2.htm [accessed 25 November 2014].

Roberts G, Myatt GJ, Johnson WP, Cross KP, Blower PE. 2000. LeadScope†: Software for Exploring Large Sets of Screening Data. J. Chem. Inf. Comput. Sci. 40:1302–1314.

Roncaglioni A, Piclin N, Pintore M, Benfenati E. 2008. Binary classification models for endocrine disrupter effects mediated through the estrogen receptor. SAR QSAR Environ. Res. 19:697–733.

Rotroff DM, Dix DJ, Houck KA, Knudsen TB, Martin MT, McLaurin KW, et al. 2013. Using in Vitro High Throughput Screening Assays to Identify Potential Endocrine-Disrupting Chemicals. Environ. Health Perspect. 121:7–14.

Royal Society of Chemistry. 2015. ChemSpider API Services. Available: http://www.chemspider.com/AboutServices.aspx [accessed 28 January 2015].

Rybacka A, Rudén C, Tetko IV, Andersson PL. 2015. Identifying potential endocrine disruptors among industrial chemicals and their metabolites – development and evaluation of in silico tools. Chemosphere 139:372–378; doi:10.1016/j.chemosphere.2015.07.036.

Sahigara F, Mansouri K, Ballabio D, Mauri A, Consonni V, Todeschini R. 2012. Comparison of Different Approaches to Define the Applicability Domain of QSAR Models. Molecules 17:4791–4810.

Schrödinger, LLC. 2011. *QikProp*. Schrödinger, LLC, New York, NY.

Shanle EK, Xu W. 2011. Endocrine Disrupting Chemicals Targeting Estrogen Receptor Signaling: Identification and Mechanisms of Action. Chem. Res. Toxicol. 24:6–19.

Shen J, Xu L, Fang H, Richard AM, Bray JD, Judson RS, et al. 2013. EADB: an estrogenic activity database for assessing potential endocrine activity. Toxicol. Sci. Off. J. Soc. Toxicol. 135:277–291.

Shukla SJ, Huang R, Austin CP, Xia M. 2010. The future of toxicity testing: a focus on in vitro methods using a quantitative high-throughput screening platform. Drug Discov. Today 15:997–1007.

Sitzmann M, Ihlenfeldt W-D, Nicklaus MC. 2010. Tautomerism in large databases. J. Comput. Aided Mol. Des. 24:521–551.

Slavov SH, Pearce BA, Buzatu DA, Wilkes JG, Beger RD. 2013. Complementary PLS and KNN algorithms for improved 3D-QSDAR consensus modeling of AhR binding. J. Cheminformatics 5:47.

Ståhle L, Wold S. 1987. Partial least squares analysis with cross-validation for the two-class problem: A Monte Carlo study. J. Chemom. 1:185–196.

Steinmetz FP, Enoch SJ, Madden JC, Nelms MD, Rodriguez-Sanchez N, Rowe PH, et al. 2014. Methods for assigning confidence to toxicity data with multiple values--Identifying experimental outliers. Sci. Total Environ. 482-483:358–365.

Sung E, Turan N, Ho PW-L, Ho S-L, Jarratt PDB, Waring RH, et al. 2012. Detection of endocrine disruptors – from simple assays to whole genome scanning. Int. J. Androl. 35:407–414.

Sushko I, Novotarskyi S, Körner R, Pandey AK, Rupp M, Teetz W, et al. 2011. Online chemical modeling environment (OCHEM): web platform for data storage, model development and publishing of chemical information. J. Comput. Aided Mol. Des. 25:533–554; doi:10.1007/s10822-011-9440-2.

Taha MO, Tarairah M, Zalloum H, Abu-Sheikha G. 2010. Pharmacophore and QSAR modeling of estrogen receptor beta ligands and subsequent validation and in silico search for new hits. J. Mol. Graph. Model. 28:383–400.

Talete srl. 2012. *DRAGON (Software for Molecular Descriptor Calculations)*. Talete srl, Milano, Italy.

Tetko IV. 2002a. Associative neural network. Neural Process. Lett. 16:187–199.

Tetko IV. 2002b. Neural network studies. 4. Introduction to associative neural networks. J. Chem. Inf. Comput. Sci. 42:717–728.

Tetko IV, Sushko Y, Novotarskyi S, Patiny L, Kondratov I, Petrenko AE, et al. 2014. How accurately can we predict the melting points of drug-like compounds? J. Chem. Inf. Model. 54:3320–3329; doi:10.1021/ci5005288.

Tice RR, Austin CP, Kavlock RJ, Bucher JR. 2013. Improving the human hazard characterization of chemicals: a Tox21 update. Environ. Health Perspect. 121:756–765; doi:10.1289/ehp.1205784.

Tong W, Hong H, Fang H, Xie Q, Perkins R. 2003. Decision forest: Combining the predictions of multiple independent decision tree models. J. Chem. Inf. Comput. Sci. 43: 525–531.

Trisciuzzi D, Dominico A, Mansouri K, Judson R, Cellamare S, Catto M, et al. 2015. Docking-based classification models for exploratory toxicology studies on high-quality estrogenic experimental data. Future Med. Chem. (In Press); doi:10.4155/FMC.15.103.

UNEP and WHO. 2013. State of the Science of Endocrine Disrupting Chemicals - 2012. Available: http://www.unep.org/chemicalsandwaste/UNEPsWork/EndocrineDisruptingChemicals/tabid/130226/Default.aspx/ [accessed 2 March 2015].

US EPA. 2014a. CPCat: Chemical and Product Categories. Available: http://actor.epa.gov/cpcat/faces/home.xhtml [accessed 26 November 2014].

US EPA. 2014b. EPI Suite Data. Available: http://esc.syrres.com/interkow/EpiSuiteData.htm [accessed 26 April 2014].

US EPA. 2015. Office of Pollution Prevention and Toxics Homepage. Available: http://www.epa.gov/oppt/ [accessed 25 November 2014].

US EPA. 1996. The Safe Drinking Water Act Amendments of 1996. Available: http://water.epa.gov/lawsregs/guidance/sdwa/theme.cfm [accessed 25 November 2014].

US EPA-NCCT. 2016. Collaborative Estrogen Receptor Activity Prediction Project Data. Available: http://www.epa.gov/chemical-research/toxicity-forecaster-toxcasttm-data [accessed 2 February 2016].

US EPA-NCCT. 2014a. DSSTox. Available: http://www.epa.gov/ncct/dsstox/ [accessed 26 November 2014].

US EPA-NCCT. 2014b. Edsp21 Dashboard. Available: http://actor.epa.gov/edsp21/ [accessed 12 January 2015].

US EPA-NCCT. 2014c. U.S. EPA’s Endocrine Disruptor Screening Program (EDSP) home page. Available: http://www.epa.gov/endo/#universe [accessed 12 January 2015].

US FDA. 1996. *Compilation of Laws Enforced by the U.S. Food and Drug Administration and Related Statutes*. U.S. Food and Drug Administration ; U.S. G.P.O, Rockville, MD : Washington, DC.

Varnek A, Fourches D, Horvath D, Klimchuk O, Gaudin C, Vayer P, et al. 2008. ISIDA - Platform for virtual screening based on fragment and pharmacophoric descriptors. Curr. Comput. Aided Drug Des. 4: 191–198.

Vedani A, Smiesko M. 2009. In silico toxicology in drug discovery - concepts based on three-dimensional models. Altern. Lab. Anim. ATLA 37: 477–496.

Wedebye EB, Niemelä JR, Nikolov NG, Dybdahl M. 2013. Use of QSAR to identify potential CMR substances of relevance under the REACH regulation. The Danish EPA. Project No. 1503. URL:http://www2.mst.dk/Udgiv/publications/2013/09/978-87-93026-48-3.pdf

Wetmore BA, Wambaugh JF, Ferguson SS, Sochaski MA, Rotroff DM, Freeman K, et al. 2012. Integration of dosimetry, exposure, and high-throughput screening data in chemical toxicity assessment. Toxicol. Sci. Off. J. Soc. Toxicol. 125:157–174.

Wold S, Sjöström M, Eriksson L. 2001. PLS-regression: a basic tool of chemometrics. Chemom. Intell. Lab. Syst. 58:109–130.

Worth AP, Bassan A, Gallegos A, Netzeva TI, Patlewicz G, Pavan M, et al. 2005. The Characterisation of (Quantitative) Structure-Activity Relationships: Preliminary Guidance. EUROPEAN COMMISSION JOINT RESEARCH CENTRE. Report: EUR 21866 EN.

Xie Q, Ratnasinghe LD, Hong H, Perkins R, Tang Z-Z, Hu N, et al. 2005. Decision forest analysis of 61 single nucleotide polymorphisms in a case-control study of esophageal cancer: A novel method. BMC Bioinformatics 6 Suppl 2:S4.

Yap CW. 2011. PaDEL-descriptor: An open source software to calculate molecular descriptors and fingerprints. J. Comput. Chem. 32:1466–1474.

Zakharov AV, Peach ML, Sitzmann M, Nicklaus MC. 2014. A new approach to radial basis function approximation and its application to QSAR. J. Chem. Inf. Model. 54:713–719.

Zang Q, Rotroff DM, Judson RS. 2013. Binary Classification of a Large Collection of Environmental Chemicals from Estrogen Receptor Assays by Quantitative Structure–Activity Relationship and Machine Learning Methods. J. Chem. Inf. Model. 53:3244–3261.

Table 1. Methods adopted by the participant groups (alphabetic order) in the modeling procedure

Model name	Calibration method	Descriptors software/type	Training Set (No. chemicals)	Predictions type
DTU	PLS/fragments	Leadscope	METI (595,481)/ ToxCast (1422)	Categorical
EPA_NCCT	GA + PLSDA	PADEL	ToxCast (1529)	Categorical
FDA_NCTR_DBB (Ng et al. 2014)	DF	Mold2	ToxCast (1677)	Categorical
FDA_NCTR_DSB	PLS	3D-SDAR	ToxCast (1019)	Categorical
ILS_EPA (Zang et al. 2013)	SVM + RF	Qikprop	ToxCast (1677)	Categorical
IRCCS_CART (Roncaglioni et al. 2008)	CART-VEGA	2D descriptors	METI (806)	Categorical
IRCCS_Ruleset	Ruleset	SMARTS	ToxCast (1529)	Categorical
JRC_Ispra (Poroikov et al. 2000)	PASS	MNA	—	Categorical
Lockheed Martin	kNN	Fingerprints	ToxCast (1677)	Categorical + Continuous
NIH_NCATS	Docking	AutoDock score	—	Categorical
NIH_NCI_GUSAR (Filimonov et al. 2009)	RBF-SCR	MNA, QNA	ToxCast (1677)	Categorical
NIH_NCI_PASS (Poroikov et al. 2000)	PASS	MNA	ToxCast (1677)	Categorical
OCHEM (OCHEM 2015)	*Consensus*	11 Descriptor types	ToxCast (1660)	Categorical + Continuous
RIFM	SVM	Fingerprints	ToxCast (1677)	Categorical
Umeå (Rybacka et al. 2015)	ASNN	DRAGON	METI + (Kuiper et al. 1997; Taha et al. 2010)	Categorical
UNC_MML	SVM+RF	DRAGON	ToxCast (120)	Categorical
UNIBA (Trisciuzzi et al. 2015)	Docking	GLIDE score	ToxCast (1677)	Categorical
UNIMIB	kNN	DRAGON + Fingerprints	ToxCast (1677)	Categorical
UNISTRA (Horvath et al. 2014)	SVM	ISIDA	ToxCast (1529)	Categorical + Continuous

Predictions type: A categorical model is one that provides an active/inactive call for each chemical, whereas a continuous model provides a prediction of the potency (in μM) for each active chemical.

Calibration methods: PLS (partial least-squares), PLS-DA (partial least-squares discriminant analysis), SVM (support vector machines), RF (random forest), DF (Decision forest), kNN (*k* nearest neighbors), ASNN (associative artificial neural networks), PASS (algorithm derived from Naïve Bayes classifier), RBF-SCR (self-consistent regression with radial basis function interpolation)

Table 2. Evaluation set for binary categorical models. Distribution of the number of active and inactive chemicals within the three different classes: binding, agonists and antagonists.

Class\activity	Active	Inactive	Total
Binding	1982	5301	7283
Agonist	350	5969	6319
Antagonist	284	6255	6539
Total	2017	7024	7522

The classification into actives and inactives is based on a consensus between the literature data sources in agreement.

Table 3. Evaluation set for quantitative models. Distribution of the number of chemicals in the five potency levels within the three different classes (binding, agonists and antagonists), classifications based on average scores.

Class\activity	Inactive	Very Weak	Weak	Moderate	Strong	Total
Binding	5042	685	894	72	77	6770
Agonist	5892	19	179	31	42	6163
Antagonist	6221	76	188	10	10	6505
Total	6892	702	916	81	93	7253

The classification of the chemicals in the five potency levels is based on the concentration responses from the literature sources in agreement.

Table 4. Confusion matrices of categorical *consensus* predictions for binding

Observed\Predicted	ToxCast Data Predicted actives	ToxCast Data Predicted inactives	Literature Evaluation Set (All: 7283) Predicted actives	Literature Evaluation Set (All: 7283) Predicted inactives
Observed actives	76	13	467	1515
Observed inactives	25	1415	268	5033

Table 5. Statistics of categorical *consensus* predictions for binding on ToxCast and literature data

Statistics\ used data	ToxCast Data	Literature Evaluation Set (All: 7283)	Literature Evaluation Set (>6 Sources: 1257)
Sensitivity	0.85	0.23	0.85
Specificity	0.98	0.95	0.97
Balanced accuracy	0.92	0.59	0.91

The literature data with more than 6 sources represents the most consistent part of the evaluation set.

Table 6. Number of chemicals reclassified after applying each one of the 4 prediction correction rules.

Rule used for each class	Rule 1 Agonist	Rule 1 Antagonist	Rule 1 Binding	Rule 2 Agonist	Rule 2 Antagonist	Rule 2 Binding	Rule 3 Agonist	Rule 3 Antagonist	Rule 3 Binding	Rule 4 binding
Number of chemicals	1288	2760	1587	217	14	344	145	161	38	966

Rule 1: chemicals that changed from inactive to active in the quantitative consensus based on the categorical *consensus*.

Rule 2: chemicals that changed from inactive to active in the categorical consensus based on the quantitative consensus.

Rule 3: chemicals that changed from active to inactive in the quantitative consensus based on the predictions of the categorical consensus.

Rule 4: chemicals that changed from inactive to active in the categorical binding consensus based on their agonist and antagonist activity in the categorical consensus.

Table 7. Confusion matrices of the modified categorical *consensus* predictions for binding

Observed\Predicted	ToxCast Data Predicted actives	ToxCast Data Predicted inactives	Literature Evaluation Set (All: 7283) Predicted actives	Literature Evaluation Set (All: 7283) Predicted inactives
Observed actives	83	6	597	1385
Observed inactives	40	1400	463	4838

Table 8. Statistics of the modified categorical *consensus* for binding predictions on ToxCast and literature data

Statistics \ used data	ToxCast Data	Literature Evaluation Set (All: 7283)	Literature Evaluation Set (>6 Sources: 1275)
Sensitivity	0.93	0.30	0.87
Specificity	0.97	0.91	0.94
Balanced accuracy	0.95	0.61	0.91

Figure legends

Figure 1. ROC curves of the categorical corrected consensus predictions for binding evaluated against different sets of the evaluation set with variable numbers of literature sources. The number of available chemicals in the evaluation set (between brackets) decreased with higher numbers of literature sources. The true and false positive rates are determined based on the number of actives in the different sets of the evaluation set. Boxes extend from the 25th to the 75th percentile, horizontal bars represent the median, whiskers indicate the 10th and 90th percentiles, and outliers are represented as points.

Figure 2. Box-plot of the positive class potency levels in the corrected quantitative *consensus* predictions for binding. The concordance between models is the fraction of the number of models that agrees on the prediction of a certain chemical.

Figure 3. Variation of the balanced accuracy of the corrected categorical consensus predictions for binding with positive concordance (agreement between models on predictions for active chemicals) threshold at different numbers of literature sources.

Figure 1.

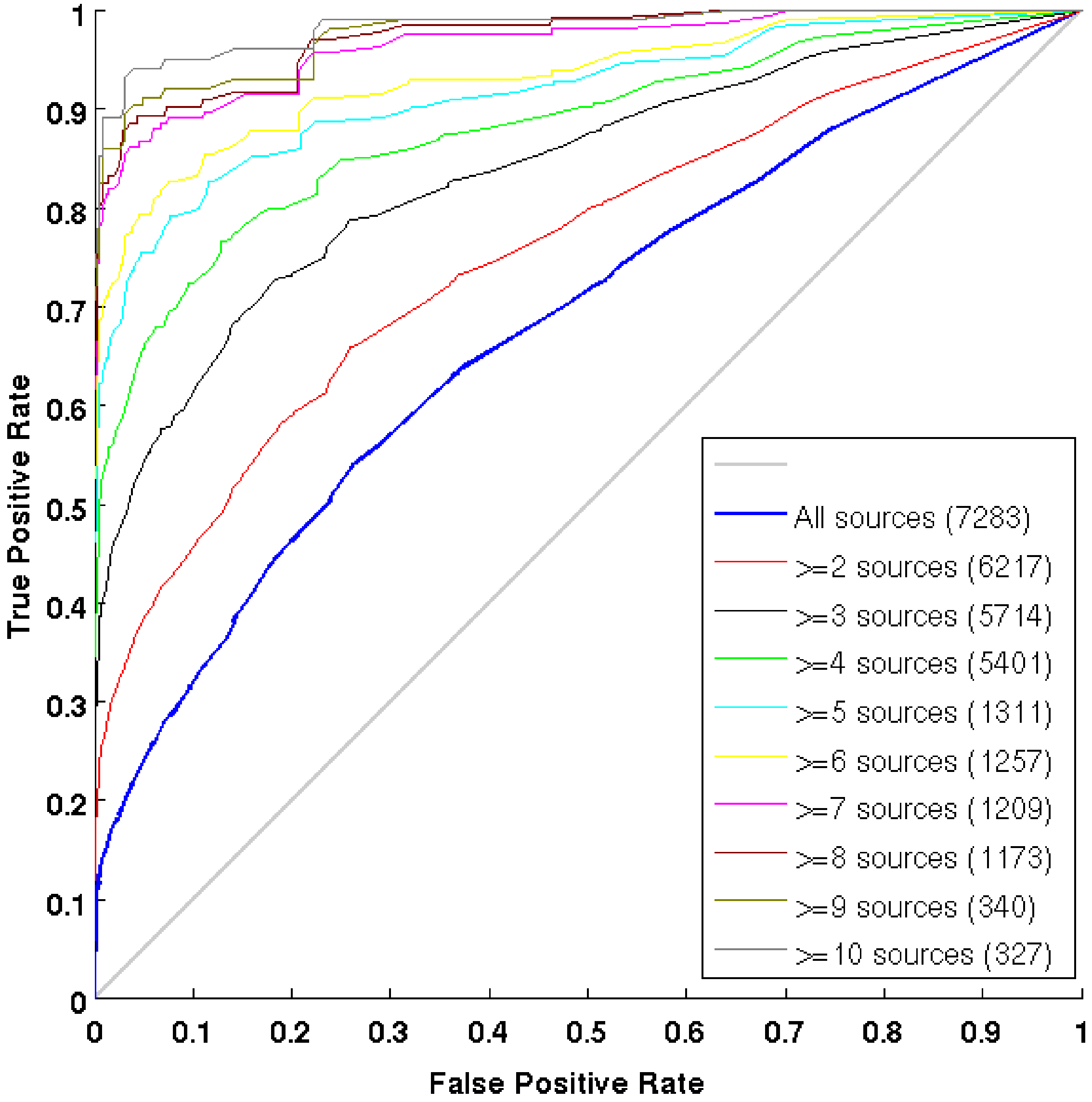

True Positive Rate
False Positive Rate
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
All sources (7283)
>=2 sources (6217)
>=3 sources (5714)
>=4 sources (5401)
>=5 sources (1311)
>=6 sources (1257)
>=7 sources (1209)
>=8 sources (1173)
>=9 sources (340)
>=10 sources (327)

Figure 2.

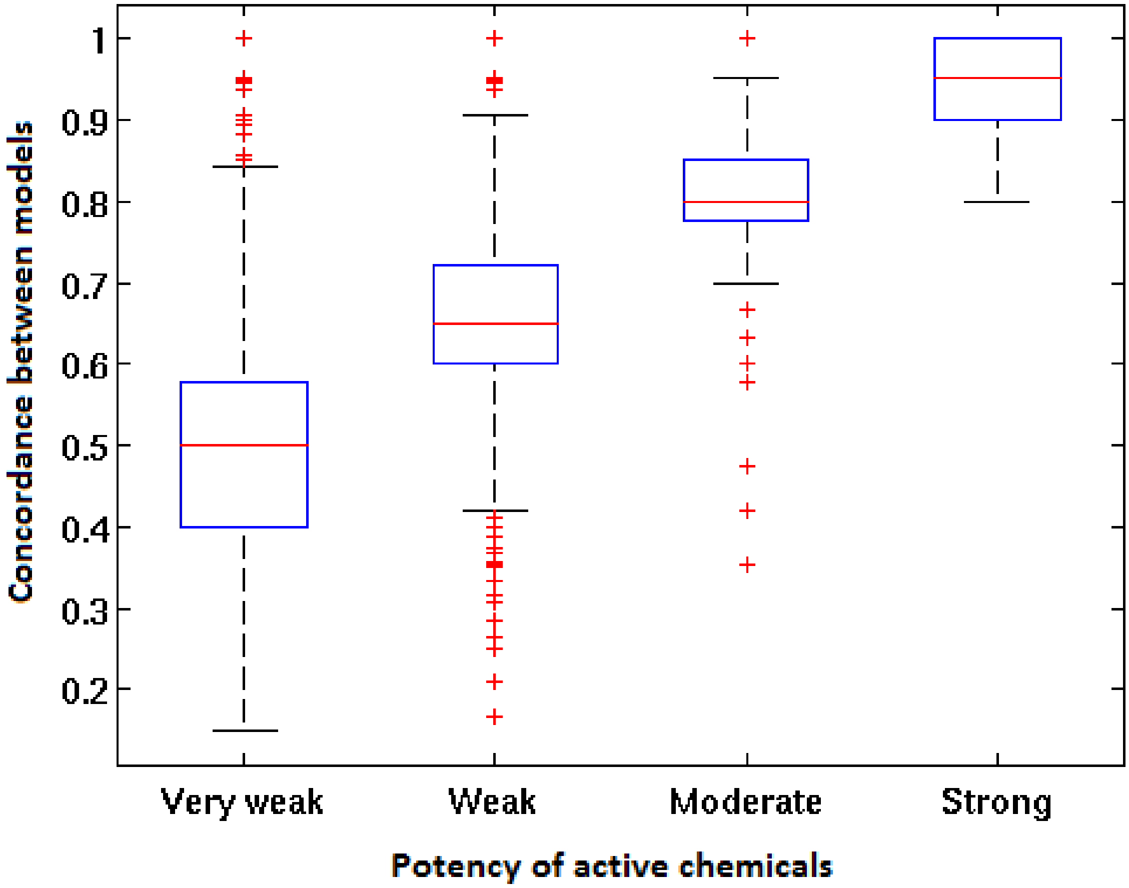
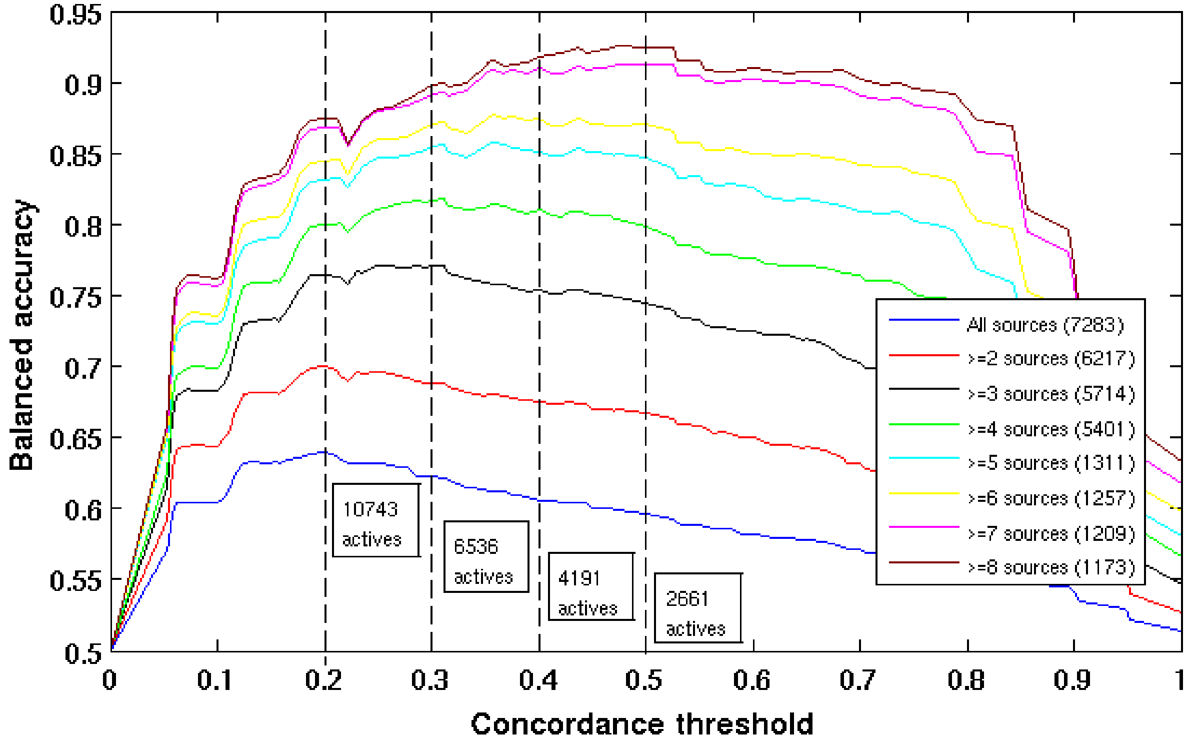

1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
Concordance between models
Very weak
Weak
Moderate
Strong
Potency of active chemicals

Figure 3.


0.95
0.9
0.85
0.8
0.75
0.7
0.65
0.6
0.55
0.5
Balanced accuracy
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
Concordance threshold
10743 actives
6536 actives
4191 actives
2661 actives
All sources (7283)
>=2 sources (6217)
>=3 sources (5714)
>=4 sources (5401)
>=5 sources (1311)
>=6 sources (1257)
>=7 sources (1209)
>=8 sources (1173)